Fairness-Aware Learning with Restriction of Universal Dependency using f-Divergences

Kazuto Fukuchi

`kazuto@mdl.cs.tsukuba.ac.jp`

Graduate School of SIE, University of Tsukuba

Jun Sakuma

`jun@cs.tsukuba.ac.jp`

Graduate School of SIE, University of Tsukuba / JST CREST

Abstract

Fairness-aware learning is a novel framework for classification tasks. Like regular empirical risk minimization (ERM), it aims to learn a classifier with a low error rate, and at the same time, for the predictions of the classifier to be independent of sensitive features, such as gender, religion, race, and ethnicity. Existing methods can achieve low dependencies on given samples, but this is not guaranteed on unseen samples. The existing fairness-aware learning algorithms employ different dependency measures, and each algorithm is specifically designed for a particular one. Such diversity makes it difficult to theoretically analyze and compare them. In this paper, we propose a general framework for fairness-aware learning that uses f-divergences and that covers most of the dependency measures employed in the existing methods. We introduce a way to estimate the f-divergences that allows us to give a unified analysis for the upper bound of the estimation error; this bound is tighter than that of the existing convergence rate analysis of the divergence estimation. With our divergence estimate, we propose a fairness-aware learning algorithm, and perform a theoretical analysis of its generalization error. Our analysis reveals that, under mild assumptions and even with enforcement of fairness, the generalization error of our method is $O(\sqrt{1/n})$, which is the same as that of the regular ERM. In addition, and more importantly, we show that, for any f-divergence, the upper bound of the estimation error of the divergence is $O(\sqrt{1/n})$. This indicates that our fairness-aware learning algorithm guarantees low dependencies on unseen samples for any dependency measure represented by an f-divergence.

1   Introduction

Recently developed information systems are being increasingly incorporating machine learning techniques for making important decisions, such as credit scoring, calculating insurance rates, and evaluating employment applications. These decisions can result in the unfair treatment, if the decisions depend on the sensitive information, such as the individual's gender, religion, race, or ethnicity. Fairness-aware learning attempts to solve this problem, and has recently received a great deal of attention [3, 16, 20]. In this paper, we consider the use of fairness-aware learning for classification problems.

Let $\mathcal{X}$ and $\mathcal{Y} = \{1, ..., c\}$ be the domain of the input and the domain of the target, respectively. In ordinary classification algorithms, the learner aims to find a hypothesis $f : \mathcal{X} \rightarrow \mathcal{Y}$ that minimizes misclassifications from a given set of iid samples $S_n = \{(x_i, y_i)\}_{i=1}^{n} \in (\mathcal{Z} = \mathcal{X} \times \mathcal{Y})^n$. In fairness-aware learning, we assume that the input x_i contains a *viewpoint* $v_i \in \mathcal{V}$, which represents the sensitive information of individuals. The learner aims to find an f that will have a low misclassification rate and for which the output of f has little dependency on the viewpoint v. For example, suppose the company want to make a hiring decision using information collected from job applicants (input x), including their age, place of residence, and work experience, but also including

their gender, religion, race, and ethnicity (viewpoint v). We wish to make hiring decisions based on the potential work performance of the job applicants (target y) via a supervised learning algorithm, $y = f(x)$. We say f is discriminatory if the output of f is dependent on the viewpoint v [15]. Fairness-aware learning attempts to avoid such unfair decisions by minimizing the dependency of the output of f on the viewpoint v [3, 6, 10, 11, 22]. Needless to say, minimization of the misclassification rate and minimization of the dependency are conflicting targets. Therefore, we need to consider the trade-off between misclassification and dependency.

The existing methods resolve this conflict by suppressing the dependency to sensitive viewpoints; this is accomplished by introducing a regularization term [6, 11, 22] or by adding constraints [3, 10, 23], to the objective function of the regular empirical risk minimization (See Section 2). Typically, such techniques can lead to predictions for which there is less dependency on the sensitive viewpoints of the given samples (empirical dependency). However, predictors with low empirical dependency do not necessarily achieve low dependency on sensitive viewpoints of unseen samples (generalization dependency). In the hiring decision example, the hypothesis f is trained with information collected from the past histories of job applicants. Predictors trained with existing methods might make fair decisions for the job applicants in the past (low empirical dependency). However, fair decisions for the job applicants in the future (low generalization dependency) are not guaranteed. Except for the method of Fukuchi and Sakuma [6], most of the existing methods have no theoretical guarantee of the generalization dependency. In [6], theoretical analysis provides a probabilistic bound on the generalization dependency, but the analysis is derived for only a specific measure of dependency.

Our contributions.

We perform a unified analysis of the fairness-aware learning with more general dependency measures based on the f-divergence [1, 4]. The f-divergence is a universal class of the divergences, which can represent most of existing divergences, including the total variational distance, the covariance, the Hellinger distance, the χ^2-divergence, and the KL-divergence. Our fairness-aware learning basically follows the framework of empirical risk minimization (ERM). The goal of fairness-aware learning is to obtain predictors with an upper bound guarantee of generalization dependency; however, it cannot be directly evaluated because the underlying distribution is not observable. We thus derive an upper bound of the generalization dependency by the empirical dependency plus two extra terms. Our framework achieves the fairness of the resultant predictors by restricting the class of hypotheses to those with low empirical dependency. Thus, the upper bound of the generalization dependency of the predictors can be theoretically derived by using the bound.

The contributions of this study are two-fold. First, we propose a novel generalized procedure for estimating the f-divergences for fairness-aware learning. Our estimation method can be regarded as a generalization of [12, 14, 18]. As already stated, we constrain the hypothesis class by the f-divergence for guarantee of fairness. It is thus important to derive a tighter upper bound of the f-divergence to achieve lower generalization dependency. Existing divergence estimation methods [14] provides an upper bound of the f-divergence; however, the bound is not suitable for our purpose for the following two reasons. First, their analysis is specifically derived for KL-divergence and cannot be expanded to the general f-divergences. Second, their bound is derived for convergence analysis, not for the upper bound of the divergence. Thus, the bound is loose for our purpose. Our generalized estimation procedure provides a tighter upper bound of the f-divergences by introducing the maximum mean discrepancy. As a result, the estimation error of the f-divergence is bounded above by the empirical maximum mean discrepancy and by $O(\sqrt{1/n})$.

Second, we formulate a general ERM framework for fairness-aware learning with employing the f-divergence. We analyze the generalization error and generalization dependency of the proposed fairness-aware learning algorithm, and we show that even when fairness is enforced, the generalization error can be bounded above by the Rademacher complexity and $O(\sqrt{1/n})$, as in the regular ERM. The generalization dependency can be bounded above by the empirical maximum mean discrepancy term and two other extra terms. Thanks to the theoretical analysis of generalization

2

dependency, we can theoretically compare the upper bound of the estimation error by dependency measures. Our analysis revealed that the divergence estimation errors for all of these divergences are $O(\sqrt{1/n})$ equally, and the Hellinger distance achieves the lowest estimation error in terms of the constant term of the probabilistic error. We also derived a convex formulation of fairness-aware learning that works with any dependency measures represented by the f-divergence. The optimization problem can be readily solved by a standard convex optimization solver.

2 Related Works

Within the setup described in the Introduction, Calders and Verwer [3] pointed out that elimination of the viewpoint from the given samples is insufficient for achieving low correlation between the output of f and the viewpoint; this is because the viewpoint has an indirect influence since it is not independent from the input. For example, when we make a hiring decision using information collected from job applicants via supervised learning, even if we train with samples that exclude race and ethnicity, the output of the resultant hypothesis f may be indirectly correlated with race or ethnicity, because the addresses of the applicants may be correlated with their race or ethnicity. Such an indirect effect is called the red-lining effect [3].

To remove the red-lining effect, existing works have attempted to construct a classifier that results in a fairer hypothesis. Calders and Verwer [3] proposed the naive Bayes classifier with fairness constraint, which employs the difference between the conditional probabilities $|\Pr(f(X) = y_+|v_+) - \Pr(f(X) = y_+|v_-)|$ where $\mathcal{Y} = \{y_+, y_-\}$ and $\mathcal{V} = \{v_+, v_-\}$. Kamiran et al. [10] and Zliobaite et al. [23] discussed various situations in which discrimination can occur, in terms of the difference of conditional probabilities. Dwork et al. [5] introduced a fairness-aware learning framework of the ERM with constraints of *statistical parity*, defined as the total variational distance between $\Pr(f(X)|v_+)$ and $\Pr(f(X)|v_-)$. Zemel et al. [22] presented an algorithm to preserve fairness in a classification setting based on statistical parity. Kamishima et al. [11] proposed a fairness-aware learning algorithm of the maximum likelihood estimation by penalizing the log-likelihood using KL-divergence between $\Pr(f(X), V)$ and $\Pr(f(X))\Pr(V)$[1]. These fairness-aware learning algorithms do not have a theoretical guarantee for the estimation error of the dependency measures. In addition, the design of these algorithms are tightly coupled with specific dependency measures. They thus have less flexibility for choosing other dependency measures.

The fairness for unseen samples can be measured by the estimation error bound of the dependency measure. Fukuchi and Sakuma [6] first derived a bound on the estimation error of a specific measure, namely the *+1/-1 neutrality risk*. They proved that the estimation error of the measure is bounded above in probability by the Rademacher complexity of the hypothesis class $\mathcal{F}$ and $O(\sqrt{1/n})$ term. Unfortunately, the analysis relies on the +1/-1 neutrality risk and cannot be generalized to other types of dependency measures.

Estimation procedures that use f-divergences that are based on iid samples have been studied extensively. For example, for the KL-divergence, method have been proposed that use nearest-neighbor distances [21] and least-squares estimations of the probability ratio [12]. To estimate f-divergences, García-García et al. [7] introduced an estimation procedure that uses loss minimization and sampling. Kanamori et al. [13] presented a divergence estimator of the f-divergences based on using the moment matching estimator [17] to estimate the probability ratio. Nguyen et al. [14] used a property of convex conjugate functions to derive the M-estimator of f-divergences, and they also derived its convergence rate. In our analysis, we derive the upper bound of the estimation error, which yields a tighter upper bound of the estimation error of dependency measures compared to the existing convergence rate analysis.

---

[1]The KL divergence between $\Pr(f(X), V)$ and $\Pr(f(X))\Pr(V)$ is known as mutual information between $f(X)$ and V.

3 Problem Formulation

Let $\mathcal{X}$ and $\mathcal{Y} = \{1, ..., c\}$ be the domain of the input and the domain of the target, respectively. We assume that the learner obtains a set of iid samples $S_n = \{(x_i, y_i)\}_{i=1}^n \in (\mathcal{Z} = \mathcal{X} \times \mathcal{Y})^n$ that are drawn from an unknown probability measure μ, which is defined on some measurable space $(\mathcal{Z}, \mathfrak{Z})$. In addition, we assume the input x_i consists of the viewpoint $v_i \in \mathcal{V}$ and various other features $w_i \in \mathcal{W}$. Thus, $\mathcal{X} = \mathcal{V} \times \mathcal{W}$. Given iid samples, the learner seeks to find a hypothesis $f : \mathcal{X} \to \mathcal{Y}$ from a class of measurable functions $\mathcal{F}$ that minimizes both the misclassification rate and the dependence on the viewpoint v. We denote (X_i, Y_i) for $i = 1, ..., n$ as the random variables of the samples, and we denote V_i as the random variable for the corresponding viewpoint.

The misclassification of the hypothesis f is evaluated by the generalization risk, which is defined as $R(f) = \mathbf{E}[\ell(Y, f(X))]$. The goal of the learner is to find the hypothesis $f^* \in \mathcal{F}$ such that

$$R(f^*) = \inf_{f \in \mathcal{F}} R(f).$$

The generalization risk $R(f)$ cannot be evaluated directly because the sample distribution μ is unknown. Instead of the generalization risk, empirical risk minimization (ERM) finds a hypothesis $f_n \in \mathcal{F}$ that minimizes the empirical risk

$$R_n(f) = \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)).$$

Minimization of the empirical risk results in a relatively low generalization risk, and the generalization risk of the resultant hypothesis converges towards that of the optimal hypothesis f^* as the number of samples increases; this has been shown theoretically [2].

3.1 A Generalized Class for Dependency Measures

For the evaluation of the dependency of the output of f on the viewpoint v, we define a general class of measures for dependency. Given that $\Pr(V)\Pr(f(X)) = \Pr(V, f(X))$ if $f(X)$ and V are statistically independent, we can evaluate this dependency by evaluating the difference between the two probability measures, $\Pr(V)\Pr(f(X))$ and $\Pr(V, f(X))$. To measure the difference between two probability measures, we use the f-divergences. Suppose $\mathbb{P}$ and $\mathbb{Q}$ are two probability measures on a compact domain $\mathcal{X}$, where $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$. The class of f-divergences, also known as the Ali–Silvey distances [1, 4], takes the form

$$D_\phi(\mathbb{P}, \mathbb{Q}) = \int \phi\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q},$$

where $\phi : \mathbb{R}^+ \to \mathbb{R}$ is a convex and lower semicontinuous function such that $\phi(1) = 0$.[2] After we define the f-divergences, we define the measure of the dependency between $f(X)$ and V as follows:

$$D_\phi(f) = D_\phi(\Pr(V)\Pr(f(X)), \Pr(V, f(X))).$$

Without loss of generality, we will assume that the subdifferential of ϕ at 1 contains 0. This can be readily confirmed by $\phi_c(u) = \phi(u) - c(u - 1)$ which does not change the value of the f-divergences $D_{\phi_c}(\mathbb{P}, \mathbb{Q}) = D_\phi(\mathbb{P}, \mathbb{Q})$ for any finite c. We will focus on the convex functions ϕ that are differentiable on $\mathbb{R}^+$ except at 1. Note that this includes most of the divergences, including the total variational distance, the Hellinger distance, the χ^2-divergence, and the KL-divergence.

---

[2]The f-divergence becomes one of the existing divergences due to the choice of ϕ; that is, it becomes the total variational distance if $\phi(u) = |u - 1|$, the Hellinger distance if $\phi(u) = (\sqrt{u} - 1)^2$, the χ^2-divergence if $\phi(u) = (u - 1)^2/u$, or the KL-divergence if $\phi(u) = (u - 1) - \ln(u)$. See Figure 1a.

3.2 Fairness-Aware Learning with Generalized Dependency Measures

In fairness-aware learning, the learner attempts to minimize both $R(f)$ and $D_\phi(f)$. However, since $\arg\min_{f\in\mathcal{F}} R(f) = \arg\min_{f\in\mathcal{F}} D_\phi(f)$ does not always hold, there exists a trade-off between $R(f)$ and $D_\phi(f)$. We thus consider a subset of $\mathcal{F}$ parameterized by $\eta \geq 0$ defined as follows:

$$\mathcal{F}_\eta = \{f \in \mathcal{F} | D_\phi(f) \leq \eta\}.$$

Thus, the goal of the fairness-aware learning is to achieve the hypothesis $f_\eta^* \in \mathcal{F}_\eta$ that satisfies

$$R(f_\eta^*) = \inf_{f\in\mathcal{F}_\eta} R(f).$$

Again, since the generalization risk cannot be evaluated directly, the learner minimizes the empirical risk as

$$\min_{f\in\mathcal{F}_\eta} R_n(f). \tag{1}$$

The objective of fairness-aware learning is to solve the optimization problem of Eq. (1). Unfortunately, $D_\phi(f)$ cannot be evaluated directly again since the underlying distribution is unobservable. In Section 4, we introduce a novel estimation procedure of the f-divergences to alleviate evaluation of $D_\phi(f)$. Then, we prove an upper bound of $D_\phi(f)$ with empirical estimation of $D_\phi(f)$ given a finite number of samples. In Section 5, the objective function of fairness-aware learning is redefined using the empirical estimation of $D_\phi(f)$.

4 Divergence Estimation

In this section, we introduce a procedure that involves minimizing the *maximum mean discrepancy* (MMD) for estimating $D_\phi(f)$, and we determine a non-asymptotic bound on the estimation error. This procedure covers the existing f-divergences or KL-divergence estimation algorithms proposed by Nguyen et al. [14], Ruderman et al. [18], and Kanamori et al. [13].

4.1 Estimation of the Divergence by Minimizing the Maximum Mean Discrepancy

To estimate $D_\phi(f)$, we first empirically estimate the probability ratio $r(V, f(X)) = d\mathrm{Pr}(V)\mathrm{Pr}(f(X))/d\mathrm{Pr}(V, f(X))$, and then we empirically evaluate $D_\phi(f)$ by using the estimated probability ratio. Since $d\mathrm{Pr}(V)\mathrm{Pr}(f(X)) = r(V, f(X))d\mathrm{Pr}(V, f(X))$ holds for the probability ratio r, the minimizer of the difference between $d\mathrm{Pr}(V)\mathrm{Pr}(f(X))$ and $r(V, f(X))d\mathrm{Pr}(V, f(X))$ is expected to be close to the probability ratio. As a measure of the disparity of $d\mathrm{Pr}(V)\mathrm{Pr}(f(X))$ and $r(V, f(X))d\mathrm{Pr}(V, f(X))$, we use the maximum mean discrepancy. Let $\mathcal{G}$ be a set of functions $g : \mathcal{V} \times \mathcal{Y} \to \mathbb{R}$. Let X' be an independent copy of X, and let V' be the viewpoint of X'. Then, the MMD with $\mathcal{G}$ between $d\mathrm{Pr}(V)\mathrm{Pr}(f(X))$ and $r(V, f(X))d\mathrm{Pr}(V, f(X))$ is defined as

$$D_{\mathrm{MMD},f}(r) = \sup_{g\in\mathcal{G}} \left[\int g(V, f(X))d(\mathrm{Pr}(V)\mathrm{Pr}(f(X))) - \int g(V, f(X))r(V, f(X))d\mathrm{Pr}(V, f(X)) \right]$$
$$= \sup_{g\in\mathcal{G}} [\mathbf{E}[g(V', f(X))] - \mathbf{E}[r(V, f(X))g(V, f(X))]]. \tag{2}$$

If r is equivalent to the probability ratio, we have $D_{\mathrm{MMD},f}(r) = 0$. However, $D_{\mathrm{MMD},f}(r) = 0 \implies d\mathrm{Pr}(V)\mathrm{Pr}(f(X)) = r(V, f(X))d\mathrm{Pr}(V, f(X))$ does always not satisfy, which requires that $\mathcal{G}$ is a set of functions on a universal kernel [19]. Therefore, the evaluation ability of the discrepancy of

MMD is dependent on the choice of $\mathcal{G}$. The U-statistics [8] gives an unbiased estimator of Eq. (2) as

$$D_{\mathrm{MMD},f,n}(r) = \sup_{g \in \mathcal{G}} \left[\frac{1}{n(n-1)} \sum_{1 \le i \neq j \le n} [g(V_i, f(X_j)) - r(V_i, f(X_i))g(V_i, f(X_i))] \right].$$

The estimator of the probability ratio r_n is obtained by minimizing $D_{\mathrm{MMD},f,n}(r)$.[3] We can add the regularizer term $\Omega(r)$ to the empirical MMD to ensure the consistency of the estimator:

$$\min_{r_n \ge 0} \; D_{\mathrm{MMD},f,n}(r_n) + \Omega(r_n). \tag{3}$$

After obtained r_n by solving Eq. (3), the f-divergence is empirically evaluated as

$$D_{\phi,n}(f) = \frac{1}{n} \sum_{i=1}^{n} \phi(r_n(V_i, f(X_i))).$$

The estimation procedure is equivalent to [14] if $\Omega(r) = \lambda_n D_{\phi,n}(f)$ where λ_n is regularizer parameter. In addition to the regularizer term, if we add the constraint $\frac{1}{n} \sum_{i=1}^{n} r_n(V_i, f(X_i)) = 1$ into Eq. (3), the estimation procedure becomes same as [18]. Letting $\Omega(r) = 0$ and the appropriate choice of $\mathcal{G}$ yields the estimation procedure of [13].

4.2 Analysis of Estimation Error

In this subsection, we show the upper bound on the estimation error

$$D_\phi(f) - D_{\phi,n}(f).$$

Surprisingly, the upper bound of the estimation error does not depend on the complexity of the class of functions $\mathcal{G}$. In what follows, we use $D_\phi^r(f) = \mathbf{E}[\phi(r(V, f(X)))]$ and $D_{\phi,n}^r(f) = \sum_{i=1}^{n} \phi(r(V_i, f(X_i)))/n$. In addition, we denotes the true probability ratio as $r^*(V, f(X))$.

The following theorem states the probabilistic upper bound on the estimation error.

Theorem 1. *Let r_n be the probability ratio estimated from the obtained set of samples S_n. Suppose that the class of the functions $\mathcal{G}$ of the MMD contains $\partial \phi(r^*(V, f(X)))/a$, where $\partial \phi$ is an element of the subdifferential of ϕ, $a > 0$ is some constant, and $c_\ell \le r^*(V, f(X)) \le c_u$ almost surely, where $c_\ell \in (0, 1]$ and $c_u \in [1, \infty)$. Then, with probability at least $1 - e^{-t}$*

$$D_\phi(f) \le D_{\phi,n}(f) + a D_{\mathrm{MMD},f,n}(r_n) + c\sqrt{\frac{2t}{n}},$$

where $c = 2\max\{\phi(c_\ell), \phi(c_u)\} + \partial\phi(c_u)c_u + \partial\phi(c_u) - 2\partial\phi(c_\ell)$.

The proof of this theorem is found in appendix C. As proved in Theorem 1, the f-divergences can be bounded above by the empirical f-divergences, the empirical MMD, and $O(\sqrt{1/n})$. We minimize the error between the f-divergences and the empirical f-divergences by minimizing the empirical MMD. In addition, the error bound does not depend on the complexity of $\mathcal{G}$. This implies that in order to guarantee the upper bound on the f-divergences, we should choose $\mathcal{G}$ so that it is large enough to satisfy $\partial\phi(r^*(V, f(X))) \in \mathcal{G}$. A large $\mathcal{G}$, however, can lead an over estimation of the f-divergences.

The convergence rate, i.e., the absolute value of the estimation error, as shown by [14] is dependent on the convergence rate of the empirical process with respect to $\mathcal{G}$. However, the upper bound proved by Theorem 1 does not contain the complexity term of $\mathcal{G}$, such as Rademacher complexity, the covering entropy and the bracketing entropy, and thus is tighter than the convergence rate.

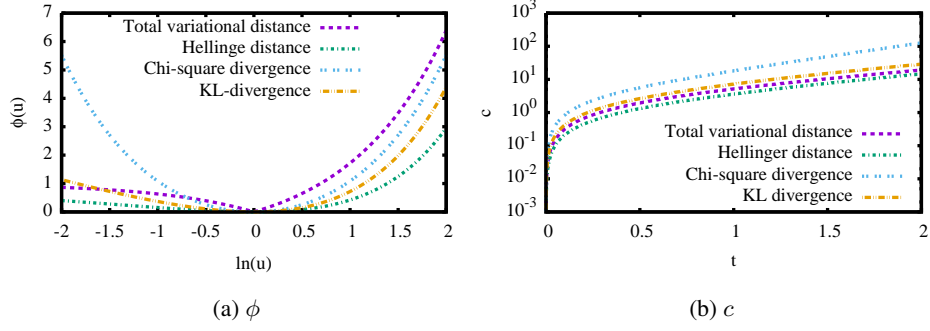[3]The efficient computation of the empirical MMD is shown in appendix A.

Figure 1: The shape of ϕ and the value of c in Theorem 1 for various ϕ, where $c_\ell = e^{-t}$ and $c_u = e^t$.

5 Fairness-Aware Learning with a Divergence Estimation

In this section, we provide an algorithm for solving Eq. (1) that includes the introduced estimation procedure for the f-divergences. We will then show that the algorithm can be formulated as a convex optimization problem.

5.1 Algorithm for Fairness-Aware Learning with f-Divergence Estimation

Following the estimation procedure described in Section 4, we define the optimization problem of our fairness-aware learning as

$$\min_{f \in \mathcal{F}, r_n \geq 0} R_n(f) \text{ sub to } D_{\phi,n}(f) + a_n D_{\mathrm{MMD},f,n}(r_n) \leq \eta, \tag{4}$$

where a_n is the constant larger than a that was defined in Theorem 1. As indicated by Theorem 1, $D_\phi(f) \leq D_{\phi,n}(f) + a_n D_{\mathrm{MMD},f,n}(r_n) + O(\sqrt{1/n})$ holds, which guarantees that the f-divergence of the resultant hypothesis of Eq. (4) is less than $\eta + O(\sqrt{1/n})$.

Let us consider the effect of the choice of ϕ on the estimation error of the divergence. The upper bound of the estimation error shown in Theorem 1 does not depend on the choice of ϕ. Nevertheless, the choice of ϕ changes the constant c. Letting $c_\ell = e^{-t}$ and $c_u = e^t$, Figure 1 shows the shape of ϕ and the value of c corresponding to t for various functions ϕ. As shown in Figure 1b, the smallest c is that for the Hellinger distance, and thus of these four divergences, it has the tightest bound on the probability ratio r^*.

5.2 Optimization

The necessary condition of convexity of Eq. (4) is the linearity of the functions $g \in \mathcal{G}$ with respect to f. With mild assumptions, it can be made convex for any choice of $\mathcal{G}$ by a simple reformulation.

Assumption 1. The hypothesis $f \in \mathcal{F}$ is formed as $f(x) = \arg\max_{y \in \mathcal{Y}} \theta(x, y)$, and $\theta \in \Theta$ is linear with respect to the parameters.

With this assumption, the function on the RKHS $\mathcal{H}$ is given as $\theta(x, y) = \langle \Phi(x, y), \boldsymbol{w} \rangle_{\mathcal{H}}$, where $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathcal{H}$ and $\boldsymbol{w} \in \mathcal{H}$. The optimization problem in Eq. (4) can be rearranged as

$$\min_{f \in \mathcal{F}} R_n(f) \text{ sub to } \min_{r_n \geq 0}(D_{\phi,n}(f) + a_n D_{\mathrm{MMD},f,n}(r_n)) \leq \eta. \tag{5}$$

Since r_n is not appeared in the objective function in original optimization problem Eq. (4), we change the optimization problem so that the optimization with respect to r_n is only appeared in the constraint. Following the derivation of the dual problem in [14], we have the dual form of the constraint as

$$\min_{r_n \geq 0}(D_{\phi,n}(f) + a_n D_{\mathrm{MMD},f,n}(r_n)) = \max_{g \in \mathcal{G}}(E_{\phi,f,n}(g)),$$

where

$$E_{\phi,n}(g) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} g(v_j, f(x_i)) - \frac{1}{n}\sum_{i=1}^{n}\phi^*(g(v_i, f(x_i))) + \frac{1}{2a_n}\|g\|_{\mathcal{H}}^2,$$

and $\phi^*(v) = \sup_u(uv - \phi(u))$. Letting $\boldsymbol{\gamma} \in \mathbb{R}^{n \times c}$, we can rewrite the optimization problem in Eq. (5) as

$$\min_{f \in \mathcal{F}, \boldsymbol{\gamma} \in \mathbb{R}^{n \times c}} R_n(f) \text{ sub to } I[f(x_i) = j] = \gamma_{ij} \ \forall i, j, \ \max_{g \in \mathcal{G}} E_{\phi,n}(\boldsymbol{\gamma}, g) \leq \eta, \qquad (6)$$

where

$$E_{\phi,n}(\boldsymbol{\gamma}, g) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n}\sum_{k=1}^{c} \gamma_{ik}g(v_j, k) - \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c}\gamma_{ik}\phi^*(g(v_i, k)) + \frac{1}{2a_n}\|g\|_{\mathcal{H}}^2.$$

From the definition of f, we have $I[f(x_i) = j] = I[\max_{k \neq j}\theta(x_i, k) - \theta(x_i, j) \leq 0]$. Let $\Delta_c = \{\boldsymbol{x} \in \mathbb{R}^c | x_i \geq 0 \ \forall i, \sum_{i=1}^{c}x_i = 1\}$. Then, we relax the indicator function as follows:

$$\min_{f \in \mathcal{F}, \boldsymbol{\gamma} \in \Delta_c^n} R_n(f) \text{ sub to } \max_{k \neq j}\theta(x_i, k) - \theta(x_i, j) \leq -\gamma_{ij}, \ \max_{g \in \mathcal{G}} E_{\phi,n}(\boldsymbol{\gamma}, g) \leq \eta. \qquad (7)$$

This optimization problem is convex, and its solution is equivalent to Eq. (4). We prove this claims by the following corollary and theorem.

Corollary 1. *If Assumption 1 holds, and ℓ is convex with respect to $\theta \in \Theta$, the optimization problem in Eq. (7) is a convex optimization problem.*

Theorem 2. *The solution of the optimization problem in Eq. (7) is equivalent to the solution of Eq. (6).*

The proofs of the corollary and the theorem can be found in appendix C.

6 Generalization Error Analysis

We consider the generalization error bound of the learned hypothesis $f_n \in \mathcal{F}$ that is obtained by the algorithm described in Section 5. In our analysis, we use the two type of the *Rademacher complexity*, which measures the complexity of the class of the functions $f : \mathcal{Z} \to \mathbb{R}$ and are defined as

$$\mathcal{R}_n(\mathcal{F}) = \mathbf{E}\left[\sup_{f \in \mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\right], \quad \mathcal{R}_n^{\mathrm{abs}}(\mathcal{F}) = \mathbf{E}\left[\sup_{f \in \mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i f(Z_i)\right|\right],$$

where (σ_i) are the independent Rademacher variables, that is, $\Pr(\sigma_i = +1) = \Pr(\sigma_i = -1) = 1/2$.

In the generalization error analysis, since our fairness-aware learning algorithm have the probabilistic error $c\sqrt{2\tau/n}$, we consider the set of hypotheses defined as

$$\mathcal{F}_\tau = \{f \in \mathcal{F}|D_\phi(f) \leq \eta + c\sqrt{2\tau/n}\}.$$

where c is defined as in Theorem 1. Theorem 1 shows that with probability at least $1 - e^{-\tau}$, $f_n \in \mathcal{F}_\tau$. Hence, application of the theorem in [2], which is appeared in appendix B, yields the generalization error bound for our algorithm.

Corollary 2. *Let f_τ^* be a hypothesis such that $R(f_\tau^*) = \inf_{f \in \mathcal{F}_\tau} R(f)$. Let $\mathcal{H}_\tau = \{h : \mathcal{Y} \times \mathcal{X} \to \mathbb{R} | h(Y, X) = \ell(Y, f(X)) - \ell(Y, f^*(X)), f \in \mathcal{F}_\tau\}$, and let $f_n \in \mathcal{F}$ be a hypothesis learned from the obtained set of samples S_n. Suppose that $h(y, x) - h(y', x') \le c$ for any $h \in \mathcal{H}_\tau$, $y, y' \in \mathcal{Y}$, and $x, x' \in \mathcal{X}$. Then, with probability at least $1 - e^{-t} - e^{-\tau}$,*

$$R(f_n) - R(f_\tau^*) \le R_n(f_n) - R_n(f_\tau^*) + 4\mathcal{R}_n(\mathcal{H}_\tau) + \sigma_{\mathcal{H}_\tau}\sqrt{\frac{2t}{n}} + c\frac{4t}{3n}.$$

Since $\mathcal{H}_\tau \subseteq \mathcal{H}$, we have $\mathcal{R}_n(\mathcal{H}_\tau) \le \mathcal{R}_n(\mathcal{H})$, $\sigma_{\mathcal{H}_\tau} \le \sigma_\mathcal{H}$. Therefore, the convergence rate of the algorithm constrained by the f-divergences is lower than that of the algorithm without the constraint.

While our algorithm guarantees an upper bound on the f-divergences, it reduces the classification performance, as compared to the classifier learned by ERM. Accordingly, let us consider the generalization error of the optimal hypotheses with and without the restriction on the f-divergences:

$$R(f_\tau^*) - R(f^*). \tag{8}$$

This error represents the reduction in the classification performance caused by restricting the f-divergences. Since the error cannot be directly evaluated, we define the estimator of Eq. (8) as

$$R_n(f_n) - \inf_{f \in \mathcal{F}} R_n(f).$$

Our interest is to derive the convergence rate of this estimator. We denote $\ell \circ \mathcal{F}' = \{h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} | h(Y, X) = \ell(Y, f(X)), f \in \mathcal{F}'\}$ for any $\mathcal{F}' \subseteq \mathcal{F}$, then the following theorem shows the convergence rate of the estimator.

Theorem 3. *Suppose that $\ell(y, f(x)) - \ell(y', f(x')) \le c$ for any $f \in \mathcal{F}$ and $(y, x), (y', x') \in \mathcal{Z}$. Then, with probability at least $1 - 2e^{-t} - e^{-\tau}$,*

$$\left| R(f_\tau^*) - R(f^*) - \left(R_n(f_n) - \inf_{f \in \mathcal{F}} R_n(f) \right) \right|$$

$$\le 4\mathcal{R}_n^{\mathrm{abs}}(\ell \circ \mathcal{F}_\tau) + 4\mathcal{R}_n^{\mathrm{abs}}(\ell \circ \mathcal{F}) + (\sigma_{\ell \circ \mathcal{F}_\tau} + \sigma_{\ell \circ \mathcal{F}})\sqrt{\frac{2t}{n}} + c\frac{8t}{3n}.$$

The proof of this theorem is appeared in appendix C.

7   Conclusions

In this paper, we considered fairness-aware learning for a classification problem, with the aim of learning the classifier that returns the prediction with the lowest misclassification rate and the lowest dependence on the viewpoint. Our contributions are as follows: (1) We propose a novel generalized procedure for estimating the f-divergences for fairness-aware learning. Our generalized estimation procedure provides a tighter upper bound of the estimation error by introducing the maximum mean discrepancy. (2) We formulate a general ERM framework for fairness-aware learning algorithm that is based on the empirical estimation procedure of the f-divergences, and that can guarantee an upper bound on the generalization dependency. Furthermore, we provide an analysis of the generalization error of the proposed fairness-aware learning algorithm.

References

[1] SM Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 131–142, 1966.

[2] Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, pages 1497–1537, 2005.

[3] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, September 2010.

[4] Imre Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizität von markoffschen ketten. *Publications of the Mathematical Institute of Hungarian Academy of Sciences*, 8:85–108, 1963.

[5] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012.

[6] Kazuto Fukuchi and Jun Sakuma. Neutralized empirical risk minimization with generalization neutrality bound. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I*, pages 418–433, 2014.

[7] Dario García-García, Ulrike von Luxburg, and Raúl Santos-Rodríguez. Risk-based generalizations of f-divergences. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 417–424, 2011.

[8] W. Hoeffding. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19:293–325, 1948.

[9] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[10] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 869–874. IEEE, 2010.

[11] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *in Proceedings of the ECML/PKDD2012, Part II*, volume LNCS 7524, pages 35–50. Springer, 2012.

[12] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.

[13] Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *Information Theory, IEEE Transactions on*, 58(2):708–720, 2012.

[14] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *Information Theory, IEEE Transactions on*, 56(11):5847–5861, 2010.

[15] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Measuring discrimination in socially-sensitive decision records. In *Proceedings of the SIAM International Conference on Data Mining, SDM*, pages 581–592. SIAM, 2009.

[16] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568. ACM, 2008.

[17] Jing Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

[18] Avraham Ruderman, Darío García-garcía, James Petterson, and Mark D Reid. Tighter variational representations of f-divergences via restriction to probability measures. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 671–678, 2012.

[19] Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

[20] Latanya Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10:10–10:29, March 2013. ISSN 1542-7730. doi: 10.1145/2460276.2460278.

[21] Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation for multidimensional densities via-nearest-neighbor distances. *Information Theory, IEEE Transactions on*, 55(5):2392–2405, 2009.

[22] Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML (3)*, pages 325–333, 2013.

[23] Indre Zliobaite, Faisal Kamiran, and Toon Calders. Handling conditional discrimination. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 992–1001. IEEE, 2011.

A Maximum Mean Discrepancy with Functions on a Reproducing Kernel Hilbert Space

Since we need to solve the maximization problem in $D_{\mathrm{MMD},f,n}(r)$, evaluation of the empirical MMD takes considerable cost causing use of the iterative algorithm. However, if the elements of the class of the functions $\mathcal{G}$ are represented by the inner products of the parameters, which includes the functions in a reproducing kernel Hilbert space (RKHS), the empirical MMD can be efficiently calculated. Let $k : \mathcal{V} \times \mathcal{Y} \times \mathcal{V} \times \mathcal{Y} \to \mathbb{R}$ be a universal kernel, and let $\mathcal{H}$ be the RKHS induced by k. Let $\Phi : \mathcal{V} \times \mathcal{Y} \to \mathcal{H}$ be the canonical feature map induced by k.

Corollary 3. *Suppose that $\mathcal{G} = \{g | g(V, f(X)) = \langle \beta, \Phi(V, f(X)) \rangle, \|\beta\|_{\mathcal{H}} \leq 1\}$. Then, the empirical MMD is equivalent to*

$$D_{\mathrm{MMD},f,n}(r) = \left\| \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} [\Phi(V_i, f(X_j)) - r(V_i, f(X_i))\Phi(V_i, f(X_i))] \right\|_{\mathcal{H}}. \quad (9)$$

Proof of Corollary 3. From the definition of the MMD and $\mathcal{G}$, we have

$$D_{\mathrm{MMD},f,n}(r) = \sup_{\beta | \|\beta\|_{\mathcal{H}} \leq 1} \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} [\langle \beta, \Phi(V_i, f(X_j)) \rangle_{\mathcal{H}} - r(V_i, f(X_i))\langle \beta, \Phi(V_i, f(X_i)) \rangle_{\mathcal{H}}]$$

$$= \sup_{\beta | \|\beta\|_{\mathcal{H}} \leq 1} \left\langle \beta, \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} [\Phi(V_i, f(X_j)) - r(V_i, f(X_i))\Phi(V_i, f(X_i))] \right\rangle_{\mathcal{H}}.$$

Since the supremum is achieved if the direction of β is equivalent to that of $\frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} [\Phi(V_i, f(X_j)) - r(V_i, f(X_i))\Phi(V_i, f(X_i))]$, we get the claim. $\square$

For simplicity of notation, we let $\Phi(v_i, f(x_j))$ be represented by Φ_{ij}, $r(v_i, f(x_i))$ by r_i, and $k(v_i, f(x_j), v_k, f(x_\ell)) = \langle \Phi(v_i, f(x_j)), \Phi(v_k, f(x_\ell)) \rangle$ by $k_{ijk\ell}$. Then, Eq. (9) can be rearranged as

$$\frac{1}{n^2} \sum_{1 \leq i,j \leq n} r_i r_j k_{iijj} - \frac{2}{n} \sum_{i=1}^{n} r_i \left(\frac{1}{n(n-1)} \sum_{1 \leq j \neq k \leq n} k_{iijk} \right) + \frac{1}{n^2(n-1)^2} \sum_{1 \leq i \neq j, k \neq \ell \leq n} k_{ijk\ell}. \quad (10)$$

Let $\boldsymbol{Q}$ be a matrix such that $Q_{ij} = k_{iijj}$, and let $\boldsymbol{p}$ be a vector such that $p_i = \frac{1}{n-1} \sum_{i \leq j \neq k \leq n} k_{iijk}$. Let $\boldsymbol{r}$ be a vector representation of r_i. The matrix representation of the minimization of Eq. (10) is obtained as

$$\min_{\boldsymbol{r} \geq 0} \frac{1}{2} \boldsymbol{r}^T \boldsymbol{Q} \boldsymbol{r} - \boldsymbol{p}^T \boldsymbol{r}.$$

The minimizer of Eq. (10) with respect to r can be easily obtained if $\boldsymbol{Q}$ is a positive definite matrix.

11

B   Generalization Error Bound of Bartlett et al. [2]

Bartlett et al. [2] proved the following theorem for the generalization error bound based on Bousquet's inequality:

Theorem 4 (Bartlett et al. [2]). *Let $\mathcal{H} = \{h : \mathcal{Y} \times \mathcal{X} \to \mathbb{R} | h(Y, X) = \ell(Y, f(X)) - \ell(Y, f^*(X)), f \in \mathcal{F}\}$, and let $f_n \in \mathcal{F}$ be a hypothesis learned from the obtained set of samples S_n. Suppose that $h(y, x) - h(y', x') \le c$ for any $h \in \mathcal{H}$, $y, y' \in \mathcal{Y}$ and $x, x' \in \mathcal{X}$. Then, with probability at least $1 - e^{-t}$*

$$R(f_n) - R(f^*) \le R_n(f_n) - R_n(f^*) + 4\mathcal{R}_n(\mathcal{H}) + \sigma_{\mathcal{H}}\sqrt{\frac{2t}{n}} + c\frac{4t}{3n},$$

where $\sigma_{\mathcal{H}}^2 = \sup_{h \in \mathcal{H}} \mathbf{Var}[h(Y, X)]$.

C   Proofs

C.1   Proof of Theorem 1

In order to prove Theorem 1, we prove following lemmas.

Lemma 1. *Suppose that $c_\ell \le r^*(V, f(X)) \le c_u$ almost surely where $c_\ell \in (0, 1]$ and $c_u \in [1, \infty)$, then*

$$\partial\phi(c_\ell) - \partial\phi(c_u) \le$$
$$\frac{1}{2}\Big(\partial\phi(r^*(V, f(X)))r^*(V, f(X)) + \partial\phi(r^*(V', f(X')))r^*(V', f(X'))$$
$$- \partial\phi(r^*(V', f(X))) - \partial\phi(r^*(V, f(X')))\Big)$$
$$\le \partial\phi(c_u)c_u - \partial\phi(c_\ell) \ a.s..$$

Proof. Since $\partial\phi$ is non-decreasing function due to the convexity of ϕ, we have

$$\partial\phi(c_\ell) \le \partial\phi(r^*(V', f(X))), \partial\phi(r^*(V, f(X))) \le \partial\phi(c_u) \ a.s.. \tag{11}$$

From the assumption that the subdifferential of ϕ contains zero, $\partial\phi(u) \le 0$ for $u \in (0, 1]$ and $\partial\phi(u) \ge 0$ for $u \in [1, \infty)$ which results that $\partial\phi(c_u)$ and $\partial\phi(c_\ell)$ are positive and negative, respectively. By this fact and Eq. (11), we have

$$\partial\phi(c_\ell) \le \partial\phi(r^*(V, f(X)))r^*(V, f(X)) \le \partial\phi(c_u)c_u \ a.s.. \tag{12}$$

Combining Eqs. (11) and (12) gives the claim.   □

Lemma 2. *Suppose that $c_\ell \le r^*(V, f(X)) \le c_u$ almost surely where $c_\ell \in (0, 1]$ and $c_u \in [1, \infty)$, then*

$$-\max\{\phi(c_\ell), \phi(c_u)\} \le \mathbf{E}[\phi(r^*(V, f(X)))] - \phi(r^*(V, f(X))) \le \max\{\phi(c_\ell), \phi(c_u)\} \ a.s..$$

Proof. From the assumption that the subdifferential of ϕ contains zero, $\partial\phi(u) \le 0$ for $u \in (0, 1]$ and $\partial\phi(u) \ge 0$ for $u \in [1, \infty)$ which yields that $\phi(u)$ is non-increasing in $(0, 1]$ and non-decreasing in $[1, \infty)$. Therefore, $0 \le \phi(r^*(V, f(X))) \le \max\{\phi(c_u), \phi(c_\ell)\} \ a.s.$, which gives the claim.   □

Proof of Theorem 1. The error $D_\phi(f) - D_{\phi,n}(f)$ is decomposed as

$$D_\phi(f) - D_{\phi,n}(f) = D_\phi^{r^*}(f) - D_{\phi,n}^{r^*}(f) + D_{\phi,n}^{r^*}(f) - D_{\phi,n}^{r_n}(f).$$

From the definition of the subdifferential, we have

$$
\begin{aligned}
&D_{\phi,n}^{r^*}(f) - D_{\phi,n}^{r_n}(f) \\
=&\frac{1}{n}\sum_{i=1}^{n}[\phi(r^*(V_i, f(X_i))) - \phi(r_n(V_i, f(X_i)))] \\
\leq&\frac{1}{n}\sum_{i=1}^{n}\partial\phi(r^*(V_i, f(X_i)))(r^*(V_i, f(X_i)) - r_n(V_i, f(X_i))) \\
=&\frac{1}{n}\sum_{i=1}^{n}\partial\phi(r^*(V_i, f(X_i)))r^*(V_i, f(X_i)) - \frac{1}{n}\sum_{i=1}^{n}\partial\phi(r^*(V_i, f(X_i)))r_n(V_i, f(X_i)). \quad (13)
\end{aligned}
$$

Since the left term in Eq. (13) is regarded as the empirical mean of $\partial\phi(r^*(V_i, f(X_i)))$ with respect to $\Pr(V, f(X))$ weighted by $r^*(V, f(X)) = d\Pr(V)\Pr(f(X))/d\Pr(V, f(X))$, it is well approximated by the U-statistics of the expectation of $\partial\phi(r^*(V_i, f(X_i)))$ with respect to $\Pr(V)\Pr(f(X))$. We thus decompose the right hand side in Eq. (13) using the U-statistics of $\partial\phi(r^*(V_i, f(X_i)))$ as

$$
\begin{aligned}
&D_{\phi,n}^{r^*}(f) - D_{\phi,n}^{r_n}(f) \\
\leq&\frac{1}{n}\sum_{i=1}^{n}\partial\phi(r^*(V_i, f(X_i)))r^*(V_i, f(X_i)) - \frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n}\partial\phi(r^*(V_i, f(X_j))) \\
&+ \frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n}[\partial\phi(r^*(V_i, f(X_j))) - r_n(V_i, f(X_i))\partial\phi(r^*(V_i, f(X_i)))]. \quad (14)
\end{aligned}
$$

Since $\mathcal{G}$ contains $\partial\phi(r^*(V, f(X)))/a$, the last term in Eq. (14) is bounded above by the MMD as

$$
\begin{aligned}
&\frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n}[\partial\phi(r^*(V_i, f(X_j))) - r_n(V_i, f(X_i))\partial\phi(r^*(V_i, f(X_i)))] \\
\leq&a\sup_{g\in\mathcal{G}}\frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n}[g(V_i, f(X_j)) - r_n(V_i, f(X_i))g(V_i, f(X_i))] \\
=&aD_{\mathrm{MMD},f,n}(r_n).
\end{aligned}
$$

Letting the first two terms in Eq. (14) be

$$
U_{\phi,n}(f) = \frac{1}{n}\sum_{i=1}^{n}\partial\phi(r^*(V_i, f(X_i)))r^*(V_i, f(X_i)) - \frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n}\partial\phi(r^*(V_i, f(X_j))),
$$

the error is bounded above as

$$
D_{\phi}(f) - D_{\phi,n}(f) \leq D_{\phi}^{r^*}(f) - D_{\phi,n}^{r^*}(f) + U_{\phi,n}(f) + aD_{\mathrm{MMD},f,n}(r_n).
$$

Next, we derive the probabilistic bound on the $D_{\phi}^{r^*}(f) - D_{\phi,n}^{r^*}(f) + U_{\phi,n}(f)$. The expectations of $U_{\phi,n}(f)$ is equivalent to zero

$$
\begin{aligned}
\mathbf{E}[U_{\phi,n}(f)] =&\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}\partial\phi(r^*(V_i, f(X_i)))r^*(V_i, f(X_i)) - \frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n}\partial\phi(r^*(V_i, f(X_j)))\right] \\
=&\mathbf{E}[\partial\phi(r^*(V, f(X)))r^*(V, f(X))] - \mathbf{E}[\partial\phi(r^*(V', f(X)))] \\
=&\mathbf{E}[\partial\phi(r^*(V', f(X)))] - \mathbf{E}[\partial\phi(r^*(V', f(X)))] \\
=&0. \quad (15)
\end{aligned}
$$

As proved the almost surely bound in Lemmas 1 and 2, application of the exponential inequality for the U-statistics [9] gives with probability at least $1 - e^{-t}$

$$
D_{\phi}^{r^*}(f) - D_{\phi,n}^{r^*}(f) + U_{\phi,n}(f) \leq \mathbf{E}[D_{\phi}^{r^*}(f) - D_{\phi,n}^{r^*}(f) + U_{\phi,n}(f)] + c\sqrt{\frac{2t}{n}},
$$

where c is a constant defined as in the claim. As shown in Eq. (15), $\mathbf{E}[U_{\phi,n}(f)] = 0$. In addition, since $D_{\phi,n}^{r^*}(f)$ is the unbiased estimator of $D_\phi^{r^*}(f)$, $\mathbf{E}[D_\phi^{r^*}(f)] = D_\phi^{r^*}(f)$. $\qquad\square$

C.2 Proof of Theorem 2 and Corollary 1

Proof of Theorem 2. For any i, if $\max_{k \neq j} \theta(x_i, k) - \theta(x_i, j) \leq 0$ holds for any $j \in \mathcal{I}_i \subseteq \{1, ..., c\}$ such that $|\mathcal{I}_i| > 1$, $\max_{k \neq j} \theta(x_i, k) - \theta(x_i, j)$ is positive for all j because of the definition of max. Thus, this case violates the first constraint in Eq. (7) since $\sum_{j=1}^c \gamma_{ij} = 1$ and $\gamma_{ij} \geq 0$ require $\gamma_{ij} > 0$ for some j. Therefore, for any i and j, if $\max_{k \neq j} \theta(x_i, k) - \theta(x_i, j) \leq 0$, then $\max_{k \neq j} \theta(x_i, k) - \theta(x_i, p) > 0$ for any $p \neq j$ because of the definition of max. This indicates that if $\boldsymbol{\gamma}$ are feasible, then only one element of $\{\gamma_{ij}\}_{j=1}^c$ is one and the others are zero for each i. Since if $\max_{k \neq j} \theta(x_i, k) - \theta(x_i, j) \leq 0$ holds then $f(x_i) = j$, $I[f(x_i) = j] = \gamma_{ij}$. Thus, since the solution of Eq. (7) holds the constraints in Eq. (6), we get the claim. $\qquad\square$

Proof of Corollary 1. Since $E_{\phi,n}(\boldsymbol{\gamma}, g)$ is linear with respect to $\boldsymbol{\gamma}$, $\max_{g \in \mathcal{G}} E_{\phi,n}(\boldsymbol{\gamma}, g)$ is a convex function with respect to $\boldsymbol{\gamma}$. Since $R_n(f)$ is convex with respect to θ because of the convexity of ℓ, the objective function in Eq. (7) is convex with respect to f and $\boldsymbol{\gamma}$. In addition, $\max_{k \neq j} \theta(x_i, k) - \theta(x_i, j) + \gamma_{ij}$ is convex with respect to θ and γ_{ij}. Thus, all constraints are convex inequalities or linear equations. $\qquad\square$

C.3 Proof of Theorem 3

For the proof of Theorem 3, we prove the following theorem.

Theorem 5. *Let $f^* \in \mathcal{F}$ be a hypothesis such that $R(f^*) = \inf_{f \in \mathcal{F}} R(f)$. Suppose that $\ell(y, f(x)) - \ell(y', f(x')) \leq c$ for any $f \in \mathcal{F}$ and $(y, x), (y', x') \in \mathcal{Z}$. Then, with probability at least $1 - e^{-t}$*

$$\left| R(f^*) - \inf_{f \in \mathcal{F}} R_n(f) \right| \leq 4\mathcal{R}_n^{\text{abs}}(\ell \circ \mathcal{F}) + \sigma_{\ell \circ \mathcal{F}} \sqrt{\frac{2t}{n}} + c\frac{4t}{3n}.$$

Proof. Since $\sup_{f \in \mathcal{F}}(R(f^*) - R(f)) = \inf_{f \in \mathcal{F}} R(f) - \inf_{f \in \mathcal{F}} R(f) = 0$, we have

$$R(f^*) - \inf_{f \in \mathcal{F}} R_n(f) = \sup_{f \in \mathcal{F}}(R(f^*) - R_n(f)) \leq \sup_{f \in \mathcal{F}}(R(f^*) - R(f)) + \sup_{f \in \mathcal{F}}(R(f) - R_n(f))$$
$$= \sup_{f \in \mathcal{F}}(R(f) - R_n(f)).$$

From the definition of sup, we have

$$\sup_{f \in \mathcal{F}}(R(f^*) - R_n(f)) \geq R(f^*) - R_n(f^*).$$

Hence, we have

$$\left| \sup_{f \in \mathcal{F}}(R(f^*) - R_n(f)) \right| \leq$$

$$\max\left\{ \left| \sup_{f \in \mathcal{F}}(R(f) - R_n(f)) \right|, |R(f^*) - R_n(f^*)| \right\} \leq \sup_{f \in \mathcal{F}}|R(f) - R_n(f)|.$$

The bound on $\sup_{f \in \mathcal{F}}|R(f) - R_n(f)|$ is derived in the same manner of the proof of Theorem 4. Application of the Bousquet's inequality gives with probability at least $1 - e^{-t}$

$$\sup_{f \in \mathcal{F}}|R(f) - R_n(f)| \leq \mathbf{E}\sup_{f \in \mathcal{F}}|R(f) - R_n(f)| + \sqrt{\frac{2v}{n}} + c\frac{t}{3n},$$

14

where $v = 2\mathbf{E}\sup_{f\in\mathcal{F}}|R(f) - R_n(f)| + \sigma_{\ell\circ\mathcal{F}}^2$. By the fact that $\sqrt{u+v} \le \sqrt{u} + \sqrt{v}$ and $2\sqrt{uv} \le \alpha u + v/\alpha$ for $u, v, \alpha \ge 0$, application of the symmetrization technique yields the claim. $\square$

Corollary 4. *Let $f_\tau^* \in \mathcal{F}_\tau$ be a hypothesis such that $R(f^*) = \inf_{f\in\mathcal{F}_\tau} R(f)$. Suppose that $\ell(y, f(x)) - \ell(y', f(x')) \le c$ for any $f \in \mathcal{F}$ and $(y, x), (y', x') \in \mathcal{Z}$. Then, with probability at least $1 - e^{-t} - e^{-\tau}$*

$$|R(f_\tau^*) - R_n(f_n)| \le 4\mathcal{R}_n^{\mathrm{abs}}(\ell \circ \mathcal{F}_\tau) + \sigma_{\ell\circ\mathcal{F}_\tau}\sqrt{\frac{2t}{n}} + c\frac{4t}{3n}.$$

Proof. The proof follows the same manner of the proof of Theorem 5 expect the upper bound of $R(f^*) - R_n(f_n)$. From Theorem 1, we have with probability at least $1 - e^{-\tau}$

$$R(f^*) - R_n(f_n) \le \sup_{f\in\mathcal{F}_\tau}(R(f^*) - R_n(f)) \le \sup_{f\in\mathcal{F}_\tau}(R(f) - R_n(f)).$$

$\square$

Proof of Theorem 3. The error is bounded above as

$$\left|R(f_\tau^*) - R(f^*) - \left(R_n(f_n) - \inf_{f\in\mathcal{F}} R_n(f)\right)\right| \le |R(f_\tau^*) - R_n(f_n)| + \left|R(f^*) - \inf_{f\in\mathcal{F}} R_n(f)\right|.$$

Combining Theorem 5 and Corollary 4 gives the claim. $\square$